Annotated Corpus of Pre-Standardized Balkan Slavic Literature

Šimko, Ivan

Abstract: The corpus contains 23 linguistically annotated samples of "damaskini" and other Balkan Slavic manuscripts and print editions from the 15th-19th century, together with over 50 thousand tokens. The texts have similar prose narratives and topics, i.e. hagiographies and apocalyptic themes. The majority of the texts are various editions of the "Life of St. Petka" by Patriarch Euthymius of Tarnovo. The primary goal of the corpus is to provide data for studies of developments in both spoken and written language in the mentioned area and period, especially the features typical for the Balkan sprachbund, e.g. postponed articles, and the analytic infinitive. The corpus has been manually lemmatized and annotated with MULTEXT-East morphosyntactic descriptions developed specifically for this corpus (http://nl.ijs.si/ME/V6/msd/html/msd-bg-dam.html) as well as syntactically analyzed with dependency relations following the Universal Dependencies guidelines (up to Level 2 validation).The annotation reflects a wide spectrum of morphosyntactic features of both archaic (Church Slavonic) and innovative (Bulgarian, Macedonian) varieties. The corpus is available both in source .tsv and derived CoNLL-U formats. The .tsv format includes tokens including accentuation and interpunction, while CoNLL-U uses a diplomatized transcript, more useful for a full-text search. The author would like to thank the following institutions for providing the source texts - National Library "Sv. sv. Kiril i Metodii" in Sofia, Church Historical and Archive Institute in Sofia, Archive of the Bulgarian Academy of Sciences in Sofia, Sofia University "Kliment Oxridski", National Library "Ivan Vazov" in Plovdiv, National and University Library of Slovenia in Ljubljana, Russian State Library in Moscow, Library of Matica Srpska in Novi Sad - as well as to Olivier Winistörfer for his invaluable help in preparing the corpus. The update 1.1 includes: - 8 new texts, with over 20k tokens - more data: Cyrillic transcripts, lemma and cross-text references - detailed philological and description of selected texts - sources in plain text format

Annotated Corpus of Pre-Standardized Balkan Slavic Literature

corpus link - http://hdl.handle.net/11356/1368

version date - 4.11.2020

Ivan Šimko, University of Zurich

1. Introduction

The following document describes the properties of our corpus of pre-standardized Balkan Slavic literature. The sources for this corpus are *pre-standardized*: they can be dated in the period before the debate on the standardization of modern Bulgarian began to level the existing dialectal differences between the literatury traditions. In the present stage, the earliest sources can be dated to 16th century, the latest to the first half of the 19th century.

The linguistic category of *Balkan Slavic* encompasses roughly the Bulgarian, Macedonian and adjacent dialects in Serbia (Torlak). While this category can be understood in geographic terms, it is primarily a linguistic one: it represents varieties of Slavic, which have undergone convergent developments together with neighboring non-Slavic varieties in the area, including Albanian, Romani, Romance, modern Greek and Gagauz Turkish. Not all of the sources have been produced in the area with a majority of populace speaking a Balkan Slavic dialect. Philological metadata for each source specify approximate affiliation of authors or scribes within broader dialectal areas (e.g. West Macedonia, Northwestern Bulgaria...) or literary standards (Church Slavonic and its redactions).

Finally, the corpus includes works of *literature*. Each source can be seen as an integral product of a single scribe, but it also can be situated within a broader scope of a literary tradition, which in many cases reaches beyond the borders of a national literature or the respective period. Each source reflects not only phonetic and grammatical features of the scribe's individual language variety, but also the relation between the scribe and existing literary traditions and the local dialect: the scribe's desire to sanctify the text by using archaic grammatical features, or to democratize it by stylistics of secular storytelling and the use of popular loanwords. Thus, each source includes implicit information about its purpose: enlightenment of a non-educated auditory, preservation of a collective memory and symbolics, and so on.

The purpose of the corpus is to provide a resource basis for philologically informed quantitative studies concerning the development of Balkan Slavic languages, focusing on the features characterstic of this group within the Slavic family.

2. Background

Balkan Slavic dialects have undergone complex morphosyntactic changes in the last (at least) five centuries. For example, case inflection, which can be observed in the oldest (Old Church Slavonic, OCS) texts from the Balkan area and other Slavic languages, has been limited to the pronominal paradigms, which are also simplified to the point of a simple distinction between a nominative and a single oblique case, with some dative forms. Morphosyntactic functions formerly expressed by inflectional cases received new analytic means of expression. Synthetic possessive constructions with genitive (e.g. *duxъ otьc**a** vaš**ego*** 'the spirit of your father' in OCS; Lunt 2001:146), dative (*syni svět**u*** 'the sons of the light'; Lunt 2001:149) or adjectives (*učenici ioan**ovi*** 'the disciples of John'; Lunt 2001:146) have been replaced by constructions with clitic dative pronouns (e.g. *nai mladĭa+ **mu** sïnь* 'his youngest son', *PPS* 80v),

prepositions (e.g. *sínъ* ***na+*** *stárago crę́* 'son of the old king', *Berl.d.* 182r) or both (e.g. *támo pripádna* ***na+*** *ikona+ ta+* ***i na+*** *presfetáe bogoróditza* ' there she knelt in front of the icon of the Holy Mother of God', *NBKM 1064* 31r).

Another example: Balkan Slavic dialects developed a system of obligatory definiteness marking. The syntactic position of the definiteness marker has been fixed to the first element of the noun phrase (the postponed article), or as the first element itself (extended demonstrative pronouns like *tozi*, indefinite marker *edin*). Similar markers are attested already in the oldest OCS texts, but rather for deictic definiteness than for anaphoric relations and the definiteness of generally known objects. Instead, OCS marked anaphoric definiteness by the distinction of long and short form of adjectives (cf. §5.2.3.). Translations of Greek texts also prefer the use of long forms for the translation of Greek articles, the demonstratives translating other determiners (e.g. ***ho*** *kakos doulos* ***ekeinos*** -> *zlъ****i*** *rabъ* ***tъ*** 'that bad servant'; Mt 24:48 in *Cod.Zogr*.). Sources from the 17th century reflect a change in this practice, freely using clitic demonstratives along long orms and distal extended demonstratives (e.g. *isplьzě* ***on'zi*** *smókъ wt+ mě́sto+* ***to+*** *si* 'the dragon crawled out from his lair', *Tixon.d.*, Demina 1972:57).

The function of the mentioned archaic and innovative forms is not the same. The modern replacements for old possessive constructions can be used for marking indirect objects (as in the mentioned sentence *támo pripádna na+ ikona+ ta+ i*) as well. Different forms of articles may be used (at least in literature) to mark the syntactic position of the noun phrase (e.g. FEM.SG article in *koito gi turixa vъ patriaršeskǫ****tǫ*** *cьrkvǫ* 'who put them into the Patriarch's church', *Nedělnik* 1856:258). Thus, it would not be fully correct to describe them as merely "possessive" resp. "definiteness" markers.

Thanks to the amount of available pre-standardized literature, it is possible to observe the individual developments of these features over time. However, literary conservativism of the scribes complicates the precise dating of the changes. A possible solution is to statististically measure the frequency of particular features in various periods of time. Especially popular works with many editions throughout the centuries are very fruitful for this kind of research. Such works can be used to construct parallel corpora, which give results less distorted by the differences in genre or writer's style. Any attempt at the construction of such a corpus, however, requires the ability to unambiguously distinguish the aforementioned changes. The corpus must be able to reflect the full diachronic (as well as diatopic) variation of morphs (or, rather, their written forms) employed throughout the observed period in its annotation. As each morph carries lexical, morphological and syntactic information, the struggle is to make this annotation economic.

The preparation of such a corpus requires the following steps:

i. **Selection of a sample**. The sources are chosen from various dialectal areas and periods. The sample should represent major trends in both literature and dialects of the Balkan Slavic area.

ii. **Digital edition**. The variety of fonts and orthographic freedom of the scribes requires a character set capable of reflecting most of the characters used in the period. Some of these (e.g. spirits) can be omitted due to their mostly ornamental function. A common UTF-8 Latin-based character set is used to reflect works written in the alphabets used in our sources - Cyrillic, Greek and Latin. Most of our texts have clear word boundaries, but unlike today, they write clitics, articles and other monosyllabic words/particles together with longer words. Such complex orthographic words are transcribed as separate tokens (e.g. ᾽и᾽ѵпѵстинїéто -> *i+ u+ pustinïé+ to* 'and in the desert', *Berl.d*. 180r). The transcription can be supplemented by additional columns with original (Cyrillic) and diplomatic forms of the tokens. The diplomatic layer uses a simplified character set, removing sets of graphemes used for a single phonem (и ї й і ы -> *i ï ĭ ı y* -> *i*), as well as accentuation, interpunction, and auxiliary markers (e.g. for

orthographic words, numbers). This layer makes the lemmatization easier, und also helps us to train automatic recognition of morphological markers.

iii. **Lemmatization**. Individual literary sources have different vocabularies. The words uncommon in modern literary Bulgarian were lemmatized mostly according to the specific dictionary based on the *Tixonravov damaskin* (Demina et al. 2012), supported by Church Slavonic (Miklosich 1865, Cejtlin 1996) and dialectal Bulgarian (*Bulgarian Etymological Dictionary* 1972-2006) dictionaries. In rare cases of loanwords not reflected in the dictionaries mentioned above, new lemmas were introduced using roots from Turkish or other donor langauges. The lemmatization helps us also to cope with orthographic differences, especially in works radically adhering to phonetic principles.

iv. **Annotation**. Tags are assigned to individual words (tokens) to describe their morphological and syntactic features. The tagset has been customized to reflect not only productive features of modern Bulgarian and Macedonian, but also those of Church Slavonic and of other Balkan Slavic dialects. Morphological annotation contains a tag based on to the MultextEast (ME) standard. Syntactic annotation is based on the Universal Dependencies (UD) system. As these are the main tools for capturing the morphosyntactic variety of Balkan Slavic, we will explain this step in more detail in the following paragraphs.

3. Data Sources

The sources of the corpus are digital transcripts of various manuscripts and printed books originating (or written by authors or scribes from) the Balkan Slavic area. The sources were chosen in order to cover as much dialectal areas and periods as possible. However, an exact localization in space and time is not always possible, especially in case of older manuscripts. So far the sources are labelled by the conventional name (abbreviated) of the edition.

source	date	language variety	text	sentences	tokens
Vuković 1536	1536	Church Slavonic	*Life of St Petka*	147	2222
Tixon.d.	early 17th	Bela Slatina-Pleven	*Life of St Petka*	274	2472
Lov.d.	early 17th	Central Balkan	*Homily on Drinking*	98	704
Ljub.d.	late 17th	Kotel-Elena	*Life of St Petka*	276	2503
NBKM 328	1750	Samokov	*Legend of St Thaïs*	131	895
Temski r.	1764	Torlak	*Homily on Children*	225	2145
NBKM 1069	1776	Panagjurište	*Homily on Divination*	112	1168
NBKM 1423	1778	Paulician	various	242	2926
NBKM 370	1784	Slavenobulgarian	introduction	111	1214
PPS	1796	Vidin-Lom	*Joseph*, *St Petka*	529	3724
Berl.d.	1803	Moesian	*Life of St Petka*	454	4853
Nedělnik 1806	1806	Slavenobulgarian	*Life of St Petka*	216	2215
NBKM 1081	1821	Pirdop-Koprivštica	*Daniel's Prophecy*	145	1249
NBKM 1064	1820s	Subbalkan	*Life of St Petka*	337	3705
NBKM 728	1830s	Tetovo	*Life of St Petka*	97	686

The corpus contains multiple versions of the *Life of St Petka* (or *Parascheva*), a Church Slavonic hagiography composed by Patriarch Euthymius of Tarnovo in the late 14th century. The oldest source (Vuković 1536) contains a shortened version, which was later added to manuscripts containing a translation of homilies by Damaskēnos Stouditēs (†1577), hence called *damaskini*. Among these, the

damaskini of Tixonravov (*Tixon.d.*), Ljubljana (*Ljub.d.*), Lovech (*Lov.d.*), and Berlin (*Berl.d.*) were included to represent the four major editions of *damaskini* manuscripts in the popular language ("simple Bulgarian"), as described by Demina (1968:53-64). The late Greek-script *damaskin NBKM 1064* also belongs to this text tradition. Manuscript *NBKM 1423*, written with Latin letters, represents another, Catholic tradition of hagiographic texts. Most of the other manuscripts among our sources are also eclectical compositions of hagiographies and homilies from various authors (*NBKM 328*, *728*, *1069*, *1081*, *PPS*, *Temski r.*). Among these, the *sbornik* of Pop Punčo (*PPS*) will be released with full annotation as a whole later. The printed *Nedělnik* 1806 by Sophronius, Bishop of Vratsa (†1813) shows a similar eclectic structure as the *damaskini* too.

Although the language of the *damaskini* is also sometimes considered a "literary standard", as it combines features of various Balkan and Western dialects of Bulgaria (cf. Velčeva 2001, Mladenova 2007). However, later examples (*Berl.d.*, *NBKM 1064*) deviate much from the 17th century norms. For this reason, we classify them according to dialectal areas of the place of their approximate origin (based on Stojkov 2002). The language of the collections *NBKM 1069* and *1081* shows considerable influence of the *damaskini* language, too.

The language varieties include two quasi-standards. The first is the traditional Church Slavonic, used in Vuković 1536. The second is "Slavenobulgarian", a variety developed on the basis of Western Bulgarian dialects with many archaic features in the 18th century. Its major example is the *Slavenobulgarian Chronicle* by hieromonk Paisius of Hilandar. The corpus includes the introduction of the *Chronicle* from the transcript produced in Elena (*NBKM 370*). The manuscript *NBKM 328* by Josif Bradati likely presents an early variant of this variety. The choice is aimed at covering an as wide as possible area (also see the map 9.1. at the end of the article):

dialectal area	**source**
Balkan	*Ljub.d.*, *Lov.d.*, *NBKM 1064*, *1069*, *1081*
Macedonia	*NBKM 728*
Moesian	*Berl.d.*
Northwestern	*PPS*, *Tixon.d.*
Rup	*NBKM 1423*
Torlak	*Temski r.*
Western	*NBKM 328*
Slavenobulgarian	*NBKM 370*, *Nedělnik 1806*
Church Slavonic	Vuković 1536

The majority of the sources (*Berl.d.*, *Ljub.d.*, *NBKM 370*, *728*, *1064*, *1069*, *1081*, *1423*, *Nedělnik* 1806, *PPS*, Vuković 1536) were provided as digital copies in the respective libraries. The rest (*Lov.d.*, *Temski r.*, *Tixon.d.*) was transcribed on the basis of critical editions.

4. Data Structure

Each source is provided in a separate file. The files use the text encoding UTF-8, which is likely to stay as a standard for a while. They are provided in two formats: the first format is a simple tab-separated text file, which can be open both in table (e.g. MS Excel) and text editors (e.g. Notepad), and is easy to convert to other formats. The files are also provided in the CoNLL-U format, which is compatible with the interface of the repository.

Text tables are per default organized in 12 columns. The first column (TOKEN) contains the main transcript, including interpunction, suprasegmental markings, and distinguishing graphematic variants.

The second column (DIPLOMATIC) contains a transcript using a simplified character set, which is more easily searchable. The third column (LEMMA) provides lemmas in basic forms - nominative singulars for nouns, pronouns and adjectives, and 1st person present for verbs. In general, the basic form was chosen according to widely available Bulgarian and Macedonian dictionaries. If the word is not mentioned in them, other (Church Slavonic, Turkish; cf. above §2.iii) dictionaries were used to determine the lemma.

The following columns include codes (tags) for morphological and syntactic information about the token (POS_TAG, POS_EXT, SENT_ID, UD_ID, UD_NCY, UD_TYPE, UD_EXT). The following column (FOLIO) shows the page number at the first token of the page. The pages are usually numbered per folio on the right side, but multiple numberings (e.g. a Cyrillic one by the author, Arabic by the library) may be given as well. The following column (TRANSLATION) contains the English translation of the whole sentence. It is usually placed at the first token of the sentence. If the sentence is complex or its syntax is broken, the translation may be provided for each logical part (clause), each starting at the first token of these parts. Some of the sources adapted for critical philological editions also have an additional column with information marking end of lines in the original edition (EOL). If multiple chapters are included from a source (like in *PPS* and *NBKM 1423*), beginnings of particular chapters are marked in another column (CHUNK), which also can be used to mark narrative components of a story (e.g. for parallel comparisons between various editions).

The CoNLL-U files use the standardized structure (link). Primary morphological tag has been reflected as the "language specific" tag (XPOS). However, format-specific limitations prevent the use of all information, which is available in text tables.

5. Morphological tagset

The corpus uses a tagset based on the Multext-East system of morphosyntactic annotation (link). It is roughly based on the specifications for modern Bulgarian, Macedonian and Croatian. The tag is composed of a string of (max. 6) characters, in which the first capital letter defines the basic part-of-speech category, followed by lower-case characters for morphological attributes. For example, the word *jasna* FEM.SG 'clear' receives the tag AFSNN, containing the following information based on the ending *-a*:

A	adjective (part-of-speech)
F	gender: FEM (i.e. not **jasen* or **jasno*)
S	number: SG (not **jasni*)
N	case: NOM (not **jasnu*, **jasnyę* etc.)
N	(old) definiteness: short-form (not **jasnaa*)

Due to transitional character and diversity of the language of our sources, the set was customized to reflect both the features, which became unproductive later, as well as the ones, which are recent. The morphological tags do not reflect the function of the morphem (e.g. definiteness, possession), which is thus left for the interpretation of measurements. They are strictly form-based, reflecting the realization of the word in the annotated text. Some desinences are ambiguous: for example the noun ending *-a* can denote not only FEM.SG.NOM, but also MASC.SG.GEN/ACC in older texts, MASC.SG shortened article, MASC.DL.NOM, NEUT.PL.NOM; they can be distinguished only by lemmas or context. Thus all the tag positions after the category marker interpret the desinence together.

5.1. Nouns

The tag includes four attributes. We distinguish grammatical gender (1), number (2), case (3), and animacy (4). Gender, number, and (at least in CS) case are marked at all elements of the noun phrase. Animacy is a lexical attribute, but it may have an effect on the paradigm. The case attribute serves to classify the graphic shape of the ending; it does not necessarily denote the morphosyntactic role of the noun. Nouns are classified either according to gender and historical stem classes (*osnova*), based on pre-Slavic SG.NOM endings (cf. Lunt 2001:54), or according to the generalized paradigms for respective genders. This is usually a choice specific for a text, but it may vary between various texts within a source (e.g. in *Tixon.d.*).

5.1.1. Tagset for nouns. Bold script marks options commonly met in newer (i.e. non-CS) texts.

- N-gender-number-case-animacy
- Gender:
    - M - masculine
    - F - feminine
    - N - neutral
- Number:
    - S - singular
    - P - plural
    - D - dual
- Case:
- singular:
    - **N - MASC -ъ, FEM -*a*/ь, NEUT -*e*/*o***
    - **G - MASC**/NEUT **-*a***, FEM -*e*/*i*
    - **D - MASC**/NEUT **-*u***, FEM -*e*/*i*
    - A - FEM -*u*
    - L - MASC -*e*, NEUT -*i*
    - I - MASC/NEUT -*om*/*em*, FEM -*oju*/*eju*
    - **V - MASC -*e*/-*u*, FEM -*o***
    - **O - MASC -*o*/*ija***, FEM: -ъ
- plural:
    - **N - MASC/FEM -*i*/*e*/*ove*, MASC -*ove*, NEUT -*a*/*i***
    - G - MASC -*ov*, MASC/FEM -ъ, NEUT -*ii*
    - D - MASC/NEUT -*om*/*em*, FEM -*am*
    - L - MASC/FEM/NEUT -*ex*/*ax*
    - I - MASC/FEM/NEUT -*mi*
- dual:
    - **N - MASC -*a*, FEM -*e*, NEUT -*i***
    - G - MASC/FEM/NEUT -*u*
    - D - MASC/FEM/NEUT -*ama*
- Animacy: **Y/N**

5.1.2. Differences from legacy tagsets

The Bulgarian tagset (BG) marks five attributes. The first attribute distinguishes common and proper nouns. This distinction was deemed redundant, being merely semantic one. Our corpus also contains a number of words ambiguous from this aspect (*Gospod* 'Lord [God]', *Bogorodica* 'Mother of God'). The last position, instead of marking animacy, denotes definiteness. In our set, the postponed article is usually not treated as a suffix. As long as they include the demonstrative root (see below §5.3.4.), articles are handled as separate tokens. This enables us to compare their use with short demonstrative pronouns used in Church Slavonic (CS). Only shortened articles (-*a*, dialectal -ъ, -*o*) are marked on the noun or adjective. When they are graphically identical to a case marker, they are annotated as such (e.g. in ***diavola*** ***sę prestruvaše*** 'the Devil clothed himself': NMSGY, *Nedělnik* 1856: 257).

The BG tagset distinguishes only two options for the case attribute: nominative and vocative. Additionally, the Macedonian (MK) set distinguishes an oblique case, too. Our marking of cases is based on the Croatian (HR) tagset, which is used also in sets for other Slavic languages, like Russian and Slovak. The situation in HR is comparable to an earlier stage reflected in CS literature. The option "oblique" has been adopted for situations, where the shape of the ending is different from the CS cases, as defined above.

1. shortened MASC.SG article	*-o*	*stolp**o*** 'the pillar' (*PPS*)
2. anlaut of a MASC.SG article	*-o*(*t*)	*posnik**o**+ t* 'the hermit' (*NBKM 728*)
3. borrowed agent nouns	*-ija*	*gemeꙗ**i͡a*** 'sailor' (*Ljub.d.*)
4. Bogorov's FEM.SG.OBL marker	*-ǫ*	*čistot**ǫ*** '[due to her] purity' (*Nedělnik* 1856)
5. *damaskini* FEM.SG.OBL marker	*-ь*	*dš**ь**+ta* '[she gave] her soul' (*Tixon.d*.)

5.1.3. Phonetic and graphic ambiguities (e.g. differences between *i*/*y* or *e*/*ě*) are resolved by a convention based on the development of the sounds in the approximate area of origin of the text, the writer or his orthographic school (e.g. *y* → *i*, *ě* → *e* for the majority of today's Bulgaria). Ambiguous case markers are marked conventionally according to their final shape, even if the ending would be syntactically interpreted otherwise. The convention is based on the traditional order of cases in paradigms (NOM-GEN-DAT-ACC-LOC-INST). Locative is marked only if its realization would be different from the dative (e.g. *v'*+ ***životě*** *svoémъ*: NMSLN; but *na* ***zemli*** 'on Earth': NFSDN; *Nedělnik* 1806). The short forms for FEM.SG.GEN and DAT are distinguished according to the stem. Old *a*-stems (ending in a hard consonant) have *-i* (< OCS *-y*) as their SG.GEN ending (e.g. ***pet'ki*** *sláva* 'glory of Petka': NFSGY; *Tixon.d.*). Old *ja*-stems (with a palatal consonant) have *-e* (< OCS *-ę*) here (e.g. *do* ***zemlje*** 'to the ground': NFSGN; *Tixon.d.*). If an OCS fem.sg.acc *-ǫ* would occur (if we include an older CS text like *Dobr.ev*. or *NBKM 667*), it would be marked as an *o*-case. Old FEM.SG *i*-stems ending in an *-i* are marked always as genitives. Various FEM.ACC.SG forms (*-u*, *-ь*) are handled with separate markers; both A and O forms may occur in the same text. Vocatives (which may be graphically indistinguishable from locatives) are marked as such, if their function is clear from the context.

5.1.4. Old *a*-stem MASC nouns (e.g. *bašta* 'father') are handled as FEM nouns.

5.1.5. Marking of dual forms follows gender-specific rules. Some originally dual forms were generalized as plurals (*kraka* 'legs'; *rъcě* 'hands'; *oči* 'eyes'), appearing alongside plural forms of adjectives and pronouns in noun phrases. The MASC ending *-a* is handled as a dual in count forms (cf. Mirčev 1978:195; Maslov 1981:149), i.e. when numerals are included in the noun phrase (e.g. *četèri* ***dě́la*** 'four parts': NMDNN; *NBKM 1081*).

5.1.6. Adverbs based on old instrumentals are also tagged according to the shape of the ending. Adverbs with the full ending are tagged as nouns, if the ending is equal to one present in the paradigm of the respective gender and stem class (e.g. *tičiš'kom* 'running': NNSIN; *Berl.d.*). If the adverbial ending is different from any in the paradigm, the token is tagged as an adverb (e.g. *noštjá* 'in the night': R; *NBKM 1064*).

5.2. Adjectives

Similarly as nouns, adjectives distinguish four attributes. The gender (1) and number (2) are in the same positions. Case (3) is distinguished in CS in all genders, but in "simple" Bulgarian of the *damaskini* only in MASC.SG adjectives systematically. Adjectives also vary according to complexity, the old expression of definiteness (4).

5.2.1. Tagset for adjectives. Bold script marks options commonly met in newer texts.

- A-gender-number-case-complexity
- Gender: F/M/N
- Number: S/P/D
- Case:
- shortform singular:
  N - MASC -ъ, FEM -*a*, NEUT -*e*
  G - MASC/NEUT -*a*, FEM -*e*/*i*

D - MASC/NEUT -*u*, FEM -*e*/*i*

A - FEM -*u*

V - vocative -*e*

L - MASC/NEUT -*e*

I - MASC/NEUT -*om*, FEM -*oju*/*eju*

- shortform plural:

N - MASC/FEM -*i*, NEUT -*a*

G - MASC/FEM/NEUT -*ъ*

D - MASC -*om*/*em*, FEM -*am*

A - MASC -*e*

L - MASC/NEUT -*ex*, FEM -*ax*

I - NEUT -*i*, FEM -*ami*

- longform singular:

N - MASC -*i*, FEM -*aa*, NEUT -*oe*

G - MASC -*ago*, FEM -*ije*

D - MASC -*omu*, FEM -*ei*

A - FEM -*oju*/*-uju*

L - MASC -*om*

I - MASC -*im*

O - MASC -*a*/*ъ*

- longform plural:

N - MASC -*ii*, FEM -*ie*/*ee*, NEUT -*aa*

G - MASC/FEM/NEUT -*ix*

D - MASC/FEM/NEUT -*im*

A - MASC -*ee*

I - MASC/FEM -*imi*

- (old) Definiteness : Y/N

5.2.2. Differences from legacy tagsets

Adjectives are tagged differently between the respective tagsets: e.g. MK set marks gender in position (2), while BG and HR mark it on the third. The case system is comparable to the HR set, with an additional *o*-case for specific situations (cf. §5.2.5.). Animacy, marked in the HR set, is not included in our system, as it does not have an influence upon the morphology of adjectives. Degree, known from the HR and MK sets, is not marked, because the relevant markers *po* and *nai* are handled as separate tokens (while CS examples of synthetic marking are too rare in our corpus).

Old definiteness or complexity is marked as "Definiteness" in the code, basing on the HR set, although it is not equal: it denotes the difference between short and long forms only. Articles are tagged as separate tokens.

5.2.3. In standard Bulgarian, the short form is prevalent throughout the paradigm; besides few lexicalized exceptions (e.g. MASC.SG *vtori* 'second'), only MASC.SG adjectives with an article follow a long form (e.g. MASC.SG.DEF *visokijat* 'the high'). The definiteness marking by the variation of short and long forms seems to be limited to CS and Serbian texts (Lunt 2001:59, 66), but it was also reported in some Moesian dialects up to the modern times as well (Mladenov 1963). However, doubling of letters representing /i/, likely inspired by the MASC.SG.LF, can also be found on unexpected places in the texts (e.g. adverbs like *junaškyi* 'heroically'). Non-nominative short-forms are rare in the damaskini; the long forms were likely generalized before the loss of inflection. As with nouns, archaic, non- or pre-standardized variants (e.g. *visokъt*, *visokii*, *visoka*) should be all covered by combinations of the boolean longform marker and case.

5.2.4. Ambiguous shortforms are resolved in the same way as nouns, i.e. according to stem and etymology. The MASC.SG.LOC ending *-om* appearing in Serbian sources and *Nedělnik* is handled as I (e.g. *na* ***presvětlomъ*** *prestólě*: AMSIN; Vuković 1536). The irregular adjective *vsem* 'all' (< OCS MASC/NEUT.SG.LOC.SF/LF *vьsemь* or MASC/NEUT.SG.INST.LF *vьsěmь*) is resolved in the same way (as AMSIN). The one vocative form is marked only if other

elements of the noun phrase have a vocative ending, considered a short form (e.g. *fariseju* ***slěpe*** 'o blind Pharisee!': AMSVN; Lunt 2001:142).

5.2.5. The so-called *o*-cases are marked according to the rules defined for nouns. It can be used only with SG long forms: for example, *zliïo+ tъ* 'the bad' (*Dobr.ev.*) is analyzed as a long form adjective with an *o*-case ending (AMSOY) and as a separate article (PD-MSN).

5.3. Pronouns

All pronouns are tagged according to their type (1), distinguishing e.g. demonstrative and indefinite ones. Person (2) is tagged only in personal pronouns. Gender (3), number (4) and case (5) are tagged in nominal pronouns of relevant types. For economic reasons, adjectival pronouns (e.g. *svoi*, *njakakъv*) are tagged as adjectives (e.g. ***někoja*** *carica* 'a queen': AFSNY; but ***koi*** *može iskaza* 'who could tell': PQ---N; both *Tixon.d.*). The determiner function of such elements is expressed in the syntactic annotation.

5.3.1. Tagset for pronouns.

- P-type-person-gender-number-case

- Type:
  - D - demonstrative (e.g. *tazi*, *onova*, *-tъ*, *onъ*)
  - I - indefinite (e.g. *njakoi*)
  - P - personal (e.g. *az*)
  - Q - interrogative (e.g. *koi*, *što*, *kolko*)
  - R - relative (e.g. *deto*, *štoto*, *toko*, *takъvzi*)
  - X - reflexive (*sebe*, *si*, *se*)
  - Z - negative (e.g. *nikoi*)
- Person: 1/2/3
- Gender: F/M/N
- Number: S/P/D
- Case:
  - N - unmarked/general case (e.g. *az*, *ti*, *nie*, *koi*)
  - G - long G/A forms: *mene*, *tebe*, *nego*, *neja*/*nju*, *nas*, *vas*, *těx*/*nix*, *kogo*
  - D - *mi*, *ti*, *nemu*/*mu*, *nei*/*i*, *nam*, *vam*, *im*/*gim*/*xim*, *komu*
  - A - short G/A forms: *me*, *te*, *go*/*ego*, *ja*/*eę*/*gu*, *ni*, *vi*, *gi*/*ix*/*xi*
  - L - e.g. *mně*, *něm*, *nix*, *kom*
  - I - e.g. *mnoju*, *nimi*, *kym*/*sěm*
  - O - e.g. *tьzi*

5.3.2. Differences from legacy tagsets

The BG and HR sets use longer strings (up to 13 characters) to reflect various nominal and adjectival attributes at once. Our system is closer to the MK set, but it removes the clitic attribute at position (6), which is obsolete due to use of syntactic marking, as well as the definiteness attribute (7), which is only relevant for adjectival pronouns (which we handle as adjectives). The number of options for the case attribute is extended according to the HR set. The *o*-case attribute is used only for FEM.SG.OBL demonstratives (cf. §5.1.2.).

5.3.3. Conventions are set to eliminate paradigmatic and phonetic ambiguities: e.g. OCS *mene* is clearly a 1SG.GEN, but in modern Bulgarian it could also reflect 1SG.ACC. In both cases they would be annotated as PP1-SG.

5.3.4. Articles, i.e. postponed short demonstrative pronouns, are marked as separate tokens, so that the tagset is compatible with Church Slavonic sources, where their clitic character is unclear. Unlike adnominal demonstratives they occur after the word, for which they serve as a determiner. Marking of articles as separate tokens allows us to analyze their declension, as well as the (dialectally relevant) phonetic character of the hiatus vowel, emerging at the end of MASC.SG.NOM nouns carrying the articles. The position after the noun or adjective is tagged by an extension in the syntactic annotation (§6.2.2.).

5.3.5. Gender and number are conventionally not marked on indefinite, interrogative, and negative pronouns. Case can be marked, if the referent is a person (e.g. *kogò*: PQ---G; *Berl.d*.). Individual sources (*NBKM 370*, *Nedělnik* 1806 and *Temski r*.) make a difference between SG and PL.NOM (e.g. *ktò*: PQ---SN, while *koí*: PQ---PN; *Nedělnik* 1806). If the form is ambiguous (e.g. because of using unaccented *koi* for both numbers), number is excluded from the tag. Tags for demonstrative and relative pronouns, as far as analyzable as nominal pronouns, include gender and number (e.g. *koéto*: PR-NSN; *Berl.d*.) too.

5.3.6. Pronouns used in possessive constructions are handled as adjectives (e.g. *negov* 'his') or as personal pronouns (e.g. M.3SG.DAT *mu*). Possessive function is marked in the syntactic annotation as a tag extension: the pronoun is tagged as a dependent on the noun expressing the possessed object (NMOD:POSS, cf. §6.2.1.). The attributes "owner number" and "gender" (used in the HR tagset) were not included in the tagset.

5.3.7. The type of the pronoun is marked according to the attached morpheme to the main root. A pronominal root without a suffix is always marked as interrogative pronoun, even if the author uses it e.g. as a relative pronoun (e.g. ***kogi*** *po_tražíxu ʾi+ naĭdoxu čášu* 'as they searched, they found the cup'; *PPS*):

interrogative	(none)	*koga* 'when'	PQ
negative	*ni-*	***ni****koga* 'never'	PZ
indefinite	*ně-*, (*v*)*sě-*	***ně****koga* 'sometimes'	PI
relative	*-to*	*koga****to*** 'as'	PR

The following roots are tagged as pronouns: *gde*, *kak*(*o*), *kakъv* (with suffix *-to*), *koga*, *kogda*, *kolko*, *kъde*, *takъv* (with suffixes *-va* or *-zi*), *togda*, *tolko*.

5.4. Numerals

Only one attribute (type) is included in the tag. Due to their lexical and morphological traits, ordinal numerals are marked as adjectives in our system (e.g. *vtóri* 'the second one': AMSNY). If the category is ambiguous, a secondary tag can be added (e.g. PL.GEN *oboix* 'both': ML and AMPGY; count form *dvama* 'two': ML and NFSNN). As with nouns and adjectives, collective numerals and articles are marked as separate tokens (e.g. *dvama+ ta* 'the two'). Compound numerals (e.g. category 11-19: *četir+ na+ iset* '14') are split into multiple tokens. Cyrillic numerals are marked by two asterisks at the beginning and the end of the string in the corpus; the marking in the original text may vary, usually containing a titlo over the letters.

5.4.1. Tagset for numerals.

- M-Form

- Form:
  - C - alphabetic (e.g. **dı**)
  - D - digital/arabic (e.g. *14*)
  - L - letter (e.g. *četir+na+iset*)
  - R - roman (e.g. *XIV*)

5.4.2. Differences from legacy tagsets

The form attribute is used in the MK set. The "alphabetic" option has been added to mark Cyrillic numerals. In future, it might be used to mark other alphabetic numbering systems (e.g. Greek, Hebrew...) as well. Due to our handling of ordinal numerals as adjectives, other attributes were considered redundant for our system.

5.5 Verbs

We distinguish six attributes for verbs: their syntactic type (1), mood (2), tense (3), person (4), number (5) and aspect (6). The syntactic type (main/auxiliary) is a legacy from standard ME tagsets; it does not

denote any morphological feature. Attribute "mood" distinguishes finite forms (indicative, conditional) from non-finite ones (proper infinitive, participles). Tense is distinguished only in synthetic forms (aorist, imperfect, present). Negative prefix *ne* is separated even if it fuses with the stem (e.g. *nja+mam* 'I do not have'). As verbal voice is usually expressed by analytical means, it is not a part of the tag.

5.5.1. Tagset for verbs.

- V-type-mood-tense-person-number-aspect

- Type:
  - M - main
  - A - auxiliary
- Mood:
  - I - indicative
  - M - imperative
  - N - infinitive
  - O - conditional
  - P - participle
- Tense:
  - A - aorist
  - I - imperfect
  - P - present
- Person: 1/2/3
- Number: S/P
- Aspect:
  - E - perfective
  - I - imperfective

5.5.2. Differences from legacy tagsets

Our set resembles the BG and the HR, but the tag ends at position (5). Among other attributes, gender is not considered a verbal attribute (participles may be tagged as adjectives to reflect it), while voice and negation (i.e. presence of a *ne*-marker) are marked by syntactic annotation. Position (6) is newly filled with an aspect attribute (cf. below §5.5.7).

5.5.3. As with other categories, analytic constructions (future tense, infinitive) are not marked on the morphological level. Thus the tense of the main verb of the future tense is marked with P ("present"). Analytic constructions are marked at auxiliary verbs by syntactic annotation. As with the attribute "type" used for nouns and adjectives, the attribute is preserved in our system to ensure compatibility with other corpora annotated by MultextEast-based tagsets. For example, in *bgъ šte da proslavi stica+ta svoę* 'God will glorify His saint' (*Berl.d.*), the main verb *proslavi* would be tagged as VMIP3SE and the future marker *šte* as VAIP3SI.

5.5.4. In comparison to standardized Bulgarian, pre-standardized varieties also include synthetic infinitives (e.g. *nı+mózı* ***postígna*** 'cannot reach': VMN---E; *NBKM 1064*). The same tag is used also for older infinitives with the full desinence (e.g. *móžeš+li+mi* ***dáti*** 'could you give me', *PPS*).

5.5.5. Modern BG/MK use (1.) resultative (or *l*-) participles as main verbs of some compound tenses, (2.) passive past participles as main verbs with a passive meaning, and (3.) gerunds (bulg. *deepričastie*), based on old present active participles. In later texts, (4.) narrative constructions with *l*-participles built from imperfect stems occur as well. The resultative and passive suffix is extended by an adjectival ending expressing gender and number; this may be reflected by a secondary tag (only number is reflected by the verbal tag). Separate marking of the voice is unnecessary in newer varieties: it is fully expressed by the tense of the participle, for present passive/past active participles were abandoned:

1. resultative	*-l-*	*što+è tïjà* ***stórila*** 'what she did' (*Tixon.d.*)	VMP--E
2. gerund	*-ki/-št-*	***kopaïkï*** *grobo* 'digging the grave' (*NBKM 728*)	VMPP-I
3. passive	*-n-/-t-*	*što+sà* ***smirénï*** 'who are meek' (*Berl.d.*)	VMPA-E

4. narrative	*-l-*	***tečálъ** na stáę* 'he ran to the saint (*Nedělnik* 1806)	VMPI-I

5.5.6. In archaic (CS) texts, which still use both present and past forms and both voices, voice can be explicitly distinguished in the syntactic annotation as a secondary tag of the auxiliary verb (cf. below §6.2.3.2.). If the participle is used as the main predicate of the sentence, inflection can be reflected in the secondary tag (all examples from Vuković 1536):

1. PRS active	*-ę/-št-*	*skázoujušte* 'saying'	VMPP-PE	AMPNN
2. PRS passive	*-m-*	*istávajema* 'being exposed'	VMPP-SE	AFSNN
3. PST active	*-v(š)-*	*stéžav'šïi* 'having conquered'	VMPA-SI	AMSNY
4. PST passive	*-n-/-t-*	*blsvénъ* 'blessed'	VMPA-SI	AMSNN

Participles and gerunds directly dependent on nouns (e.g. *crstvujuštomu gradu* 'to the ruling city') are tagged as adjectives. A secondary tag may be mentioned to reflect the verbal attributes. If the participle serves as the main predicate of the sentence, the verbal tag may be used as the primary one. For example, *crstvujuštomu* would receive AMSDY (adjectival) as the primary tag and VMPP-SI (verbal) as the secondary one. In *što+sà smirénï*, the verbal tag is primary.

5.5.7. Aspect is handled as an inherent feature of the lemma. Previous ME tagsets used different methods: e.g. MK/UA/PL show it on the 2nd position, RU on the 9th, BG/SC do not reflect it at all. These tagsets usually list E and P as the options, with P denoting the imperfective ("progressive") aspect; this option was replaced with an I to avoid confusion. If the aspect cannot be unambiguously determined (e.g. in loanwords), the position remains empty.

5.5.8. Imperfect and aorist plural forms have likely undergone analogical levellings already in the Middle Bulgarian period (Mirčev 1978:215). Imperfect endings and aorist stems seem to have been preferred in the process. For this reason some of the modern forms, already attested in the earliest damaskini, are ambiguous (e.g. 3PL.AOR/IMPF *poslúšaxa* 'they obeyed'; *Tixon.d.*). As the generalized plural endings (*-xme*, *-xte*, *-xa*) reflect the old imperfect, tokens with them are tagged as imperfects too. Only if the token shows a clear reflex of an old aorist ending (*-xom*, *-ste*, *-šę*), it is marked as aorist (e.g. *tako **satvoriša*** 'thus they did': VMIA3PE; *NBKM 328*). Ambiguous 1SG.AOR/IMPF forms are resolved according to etymology: verbs with uncontracted IMPF stems (e.g. *ímějaxъ* 'I had': VMII1SI; Vuković 1536) are tagged as imperfects, the rest as aorists (e.g. *imax*: VMIA1SI; *Tixon.d.*).

5.6. Prepositions

Marked as "adpositions" in the tagset. The single attribute marks the nominal case or row, called by the adposition in Church Slavonic and other Slavic languages with preserved nominal inflection. By this we can distinguish e.g. *na* denoting location (tagged SL, e.g. *na zemli* 'on Earth') from those denoting direction or possession (both tagged SA, e.g. *na+ kъ́šta* 'of the house'). The choice is based on the classification by Lunt (2001:151 f.). This attribute helps us distinguish some homonymes. More recent prepositions (e.g. *spored* 'according to') have no case/row marking.

5.6.1. Tagset for prepositions.

- S-case

- Case:
  - G - e.g. *bez*, *do*, *radi*, *za* (causal)
  - D - e.g. *k*, *po* (lative), *prěmo*, *protivo*
  - A - e.g. *na*, *nad*, *za* (lative), *po* (causal)
  - L - e.g. *na*, *po*, pri, *v* (locative)
  - I - e.g. *meždu*, *nad*, *za*, (locative), *s*

5.5.2. Differences from legacy tagsets

The case marking has a precedent in the HR set, where it is used at position (3). Separate attributes for distinguishing of pre- and postpositions (usually marked at the first position), as well as between simple and compound ones, were considered redundant for our corpus. The first attribute is relevant for only two words used in CS (*radi* and *dělьma* both 'because'), the second one is inferred in the syntactic annotation.

5.7 Particles

The single attribute reflects the type, which is a lexical category. This category is generally avoided due to its ambiguous definition. For example, *da* can be classified as both a conjunction ('so that') and a particle used in analytic verbal constructions. We prefer to mark it as a conjunction (C) to avoid confusion. This is a conventional choice: its syntactic function is reflected in another layer of annotation. The verbal particle *šte* (which can still be inflected by tense in the standard Bulgarian and by person in pre-standardized varieties) is marked as an auxiliary verb (i.e. VAIP3SI, for *šteše*: VAII3SI etc.). Ambiguous forms can also be explained by a secondary tag (e.g. *e* 'this' in *NBKM 1423*: both as QD and PP3NSN).

5.7.1. Tagset for particles.

- Q-type

- Type:
  - C - comparative (e.g. *po*, *nai*)
  - D - definitive (e.g. *eto*, *že*)
  - G - general (e.g. *dori*, *makar*, *sireč*)
  - O - modal (e.g. *mai*)
  - Q - interrogative (e.g. *dali*, *li*, *nali*, *eda*)
  - Z - negative (e.g. *ne*)

5.7.2. Differences from legacy tagsets

The particles are marked similarily as in the HR set. The number of options was extended with general and comparative ones, present in the BG set. Separate attribute for distinguishing between simple and compound particles (present in the BG and MK sets) was considered redundant: it can be inferred from the syntactic annotation. Particles denoting anaphoric and deictic relations are marked as "definitive", an option adopted from the Albanian tagset.

5.8. Others

Other parts of speech have no further specifications: adverbs (R; e.g. *tamo* 'there'), conjunctions (C; e.g. *i* 'and'), interjections (I; e.g. *o* 'oh') and non-verbal elements (X; e.g. †, .:.). With such tokens, only the part-of-speech category is marked in the tag.

5.8.1. Differences from legacy tagsets

BG, HR, and MK sets use attributes "type" and "degree" for adverbs. Adverbial attribute "type" has multiple options in the MK set, but it is ambiguous: the difference between general (e.g. *dobro* 'well') and adjectival adverbs (e.g. *sigurno* 'surely') is arbitrary. Some of the types may be interpreted as other parts of speech, e.g. verbal adverbs (gerunds) are marked as participles (e.g. *kopaïkï* 'while digging': VMPP-I; *NBKM 728*). The attribute "degree" is considered redundant, because the comparative and superlative markers are handled as separate tokens.

Furthermore BG, HR and MK sets distinguish two additional attributes for conjunctions: type (coordinating or subordinating) and formation (simple and compound). As these attributes reflect rather syntactic properties, the both were removed from our tagset.

Although preserved for special cases (e.g. markers of omission), the representation of non-verbal elements as separate tokens was by rule avoided during the process of annotation. In many our sources, punctuation is not used systematically by the scribe (especially in *PPS*). For this reason, punctuation and other non-verbal symbols were attached to the preceding token. Unless analyzed as separate tokens, they are removed from the diplomatic transcripts and the CoNLL version of our files.

5.8.2. Some words are functionally ambiguous. Subordinating conjunctions, which may have either a temporal (e.g. ***egda*** *samna skaza pavelъ elika vide* 'as the dawn came, Paul explained what he saw; *NBKM 328*) or a conditional (e.g. ***egda****+ se+ sra_mueši počto+ si prišalъ* 'if you feel shame, why did you come?'; *NBKM 328*) meaning, are tagged as adverbs (R). If the root can be extended with pronominal suffixes, it is tagged as a pronoun (cf. §5.3.7.).

The following lemmas are always marked as adverbs: *dnes*, *legoma*, *malko*, *mnogo* (inflected CS forms like F.SG.INST *mnógojù* in Vuković 1536 are classed as adjectives under the lemma *mnog*), *naedno*, *ošte*, *pak*, *paki*, *sega*, *skoro*, *tamo*, *toko*, *tokmo*, *tuk*, *tъi*, *tъčiju*, *vъn*, *vъtre*, *štom*.

The following lemmas are always marked as conjunctions: *a*, *ako*, *ala*, *ali*, *ami*, *andžak*, *ašte*, *bo*, *da*, *dano*, *zaedno*, *zam*, *zatova*, *zašto* (unless context would rather denote *za+ što*), *zaštoto*, *i*, *ibo*, *ili*, *jako* (only for CS 'if'; newer *jako* 'much, very, strongly' tagged as R), *jakože*, *kato*, *nelo*, *neti*, *ni*, *niže*, *nito*, *no*, *nъ*, *obače*, *oti*, *pa*, *počto*, *poneže*, *ta*, *ubo*, *vъskuju*, *če*, *čunki*.

6. Syntactic tagset

6.1. Standard Tags

Syntax is marked on every token in three columns. The first column (UD_ID) contains a number, which denotes the position of the word in the sentence. The second column (UD_NCY) contains a number denoting the dependency of the element. Roots (sentence predicates) are tagged as 0. The third column (UD_TYPE) defines the type of the relation. The corpus uses the standard tagset defined at the Universal Dependencies website (link).

ACL - adjectival clause (root of a subordinate clause dependent on a nominal)
ADVCL - adverbial clause (root of a subordinate clause dependent on a verb)
ADVMOD - adverbial modifier (e.g. *síč'ko+to* ***naednò*** *oustávĭ* 'she left everything at once')
AMOD - adjectival modifier (e.g. ***stáę*** *pétka* 'St Petka')
APPOS - appositional modifier, modifying the immediately preceding noun (e.g. *sъs bráta+ĭ* ***evɵímïe*** 'with her brother Euthymius')
AUX - auxilla (usually 'be', 'have' or 'want' auxiliary verbs)
CASE - analytic marking of an oblique nominal (usually an adposition)
CC - coordinating conjunction (e.g. *dobrinì* ***i+zakrílinïe*** 'virtues and protection')
CONJ - conjunct (dependent on first conjunct, e.g. in *dobrinì i+zakrílinïe*, *dobrinì* will be tagged as OBJ and ***zakrílinïe*** as CONJ)
COP - copula
CSUBJ - clausal subject
DET - determiner (demonstratives, articles)
EXPL - expletives: reflexive pronouns and doubled clitica
FIXED - fixed multiple word expression (e.g. *ne+ ště́* ***da*** *trъpì* 'she wouldn't wait')
MARK - marker of subordinated clauses (usually a subordinating conjunction or relative pronoun, e.g. *slъzì* ***katò*** *wt kládenъcь istíčaxa* 'tears poured as from a well', *da*-particle in *ne+ bě´ neĭ támь* ***da****+ sa zatrìžĭ* 'there was no one there to care')
NSUBJ - main sentence subject
NMOD - nominal modifier (e.g. *xódinïe to* ***ĭ*** 'her journeys')
NUMMOD - numeric modifier (e.g. ***čétiry*** *pьprišta* '[distance of] four shots')
OBJ - direct sentence object

OBL - nominal used as an oblique argument (e.g. in *Krstítelě ïwán'na ouprïličáva sъs pósteně* 'she followed John the Baptist by fasting', *Krstítelě* would be OBJ, *pósteně* OBL, and *ïwán'na* APPOS)

ORPHAN - used in case of a head ellipsis, where simple promotion (to a head) would result in unnatural/misleading dependency relation (e.g. due to missing object *da+ iz'rečè néï, déto+ gĭ na epivatъ* ***struvaše*** 'to tell of her [deeds?], which she did in Epibates'

REPARANDUM - reparandum: stricken tokens (depend on any adjacent non-stricken element in the sentence)

ROOT - head/predicate of the sentence

VOCATIVE - vocative

6.2. Tag Extensions

Some tags can optionally be extended for more specific relations, using the fourth column (UD_EXT). Multiple extensions can be added, separated by colon <:>. When an elision causes the tagged word to be promoted to a clause head (i.e. ACL or ADVCL) or when it is marked as a conjunct (CONJ), it may receive the extension (e.g. *slъzì katò wt* ***kládenъcъ*** *istíčaxa*; *Tixon.d.*) too.

The customization of our tagsets aims to unambiguously capture features, which are expressed in one variety synthetically (e.g. by a case ending) and analytically in another one (e.g. by a clitic). In this way, the corpus can include archaic, pre-standardized and contemporary dialectal material, allowing us its comparison across space and time.

6.2.1. One set of extensions is used with nominal modifiers (NMOD) and oblique arguments (OBL), describing their relation in a more specified way (e.g. as a spatial or possessive relation). This helps us distinguish various meanings of polysemic prepositions like *na*, or to analyze the marking of syntactic functions formerly represented by nominal case endings. The tag extensions are marked in a separate column (examples from *Berl.d.*):

ABL - ablative relation (e.g. *bě´še wt+* ***mě´sto*** *epiváti* 'she was from the town of Epibates')

IOBJ - indirect object (e.g. *bъdí+* ***mĭ*** *drugár'* 'be my comrade', lit. 'be to me a comrade')

LAT - allative relation (e.g. *wtídoxa v' nbsny"* ***ográdĭ*** 'they departed to the heavenly gardens')

LOC - locative relation (e.g. *po+ síč'ky* ***svě´tь*** 'all over the world')

POSS - possessor (e.g. *sъs'+ bráta+* ***sî*** 'with her brother')

XCOM - open clausal modifier (e.g. *bъdí+ mĭ* ***drugár'*** 'be my comrade')

6.2.2. Changes in use of determiners (extended and short clitic demonstratives) can similarily be reflected by the additional extension of the DET relation. The extension can be used after tags for nouns and adjectives, if the determiner is expressed by a vowel only (e.g. *a* ***vinaro*** *crevь reče* 'but the king's cupbearer said'; *PPS*). This is an extension, which is often combined with other ones: for example, in *života čoveko* 'the life of the human' (*NBKM 1069*) both endings seem to reflect definiteness, but in *čoveko* the relation is also that of a possessor. Thus its UD tag would be extended with P_NOM:POSS, extensions are orderred alphabetically (examples from *Tixon.d.*):

EXT - an extended demonstrative (e.g. ***tázi*** *miriz'mà* 'the stench')

P_ADJ - a postponed article following an adjective (e.g. *mói+ te móšti* 'my relics')

P_NOM - a postponed article following a noun (e.g. *kov'čeg'+* ***t'+*** *sî* 'their coffin')

6.2.3. Periphrastic verbal constructions like perfect tenses or conditionals can be marked by adding the function tag to the auxiliary verb (AUX relation), markers of subjunctive relations (MARK) or their complements of fixed phrases (FIXED):

CON - conditional (e.g. *dší+te+sĭ* ***bíx'me*** *dàlĭ* 'we would give our souls away'; *Berl.d.*)

FUT - periphrastic future (e.g. ***šte+*** *da+ reče+ tája+ réčь* 'he will say the word'; *NBKM 1069*)

INF - "infinitives" after auxiliary verbs (e.g. *šte+* ***da+*** *reče+ tája+ réčь* 'he will say the word'; *NBKM 1069*)
OPT - optative (e.g. ***da*** *ne viditъ+ nas neko* 'nobody should see us'; *NBKM 238*)
PASS - passive voice (e.g. *ispálnena* ***mpésı*** 'she was filled'; *NBKM 1064*)
PRF - perfect/narrative mood (e.g. *togaī dīavolъ* ***e+****zavīdīlъ* 'then the devil was jealous'; *NBKM 728*)
PPRF - plusquamperfect (e.g. *tь´i+se* ***bě´še*** *pod_kánila* 'thus she had decided to live'; *Tixon.d.*)

6.2.3.1. The tag extension INF is applied to *da*-markers following verbs *moga* 'can' and *šta* 'want' (including variants like *xoču*, *ću* etc.), which in some dialects may complement synthetic infinitives (cf. Mirčev 1978:235), e.g. *polovina ot crstvoto si da štem* 'we would give a half of our kingdom (*Tixon.d*.). Other such verbs are *ima*: 'have' (incl. negative future marker *njama*), *načena* (cf. Lunt 2001:154), *počna*, *(v)zema* (e.g. *ze da vuni* 'it began to stink'; *NBKM 1064*) and *podbra* (*pudmpra da struva* 'she began to do [miracles]'; *NBKM 1064*) 'begin', *stiga*, *přestana* 'stop'and the negative command *nedei* 'do not' (e.g. *nedei se gnusi* 'do not be disgusted; *Tixon.d*.). The extension serves to distinguish instances, where *da*-constructions have likely replaced synthetic infinitives, from other types of subordinated clauses. It is also able to recognize phrases with multiple main verbs after a single auxiliary verb, e.g. *Ponéže níčto drúgoe táko dšu člčskoe ne+ móže da očísti i+ na pérvo_obrázie da+ privedè kakvotъ pustýnnoe i+ bezmólvnoe živénie.* 'Because nothing else is able to purify the human soul and to make it perfect so well as the silent life in a desert." (*Nedělnik* 1806).

6.2.3.2. The current design of the morphological tagset does not allow us to distinguish the voice of verbal participles (cf. §5.5.6.). Although the problem is of less relevance to non-CS texts, the extension PASS can be used with present and past participles used as roots of subordinate clauses (ACL, ADVCL).

7. Annotation Examples

7.1. Vuković 1536 (Venice, Church Slavonic, 1536), 194v

Врѣ́мениже немáлѫ ми́мошьдшоỳ. своѥ̀ ѿхóжденїе ѥ́же ѿсоýдоу разоýмѣ. 'and after some time passed, she understood that she will depart soon from this world'

text	diplomatic	lemma	PoS tag		UD tag			
Vrě´meni+	*vrěmeni*	vrěme	NNSDN		1	5	CSUBJ	
že	*že*	že	QD		2	12	CC	
ne+	*ne*	ne	QZ		3	4	AMOD	
málu	*malu*	malo	ANSDN		4	1	AMOD	
mímošьd'šoù.	*mimošъdšou*	mimoiti	VMPA-SE	ANSDN	5	12	ADVCL	
svojè	*svoje*	svoi	ANSNY		6	7	AMOD	POSS
wtxóždenïe	*otxoždenie*	otxoždenie	NNSNN		7	12	OBJ	
jé+	*je*	je	PP3NSN		8	9	NMOD	
že	*že*	že	QD		9	11	MARK	
wt+	*ot*	ot	SG		10	11	CASE	
soúdou	*soudou*	sъdě	R		11	7	ACL	OBL
razoúmě.	*razoumě*	razuměti	VMIA3SE		12	0	ROOT	

7.2. *Tixon.d.* (Bela Slatina-Pleven or Central Balkan area, early 17th century), Demina 1972:95

ȧдш̃áтатищѐ дабѫ́де накрáсныи ѝневещестьвныи рáи. 'and your soul shall be in the beautiful and immaterial Paradise'

text	diplomatic	lemma	PoS tag	UD tag		
a+	*a*	a	C	1	7	CC
dšá+	*dša*	duša	NFSNY	2	7	NSUBJ

ta+	*ta*	тъ	PD-FSN	3	2	DET	P_NOM
ti+	*ti*	ti	PP2-SD	4	2	NMOD	POSS
šte"	*šte*	šta	VAIP3SI	5	7	AUX	FUT
(da)+	*da*	da	C	6	5	FIXED	INF
búde	*bude*	бъда	VMIP3SE	7	0	ROOT	
na+	*na*	na	SL	8	13	CASE	
krásnyi	*krasnii*	krasen	AMSNY	9	13	AMOD	
i+	*i*	i	C	10	12	CC	
ne+	*ne*	ne	QZ	11	12	AMOD	
veštestь´vnyi	*veštestъvnii*	veštestven	AMSNY	12	9	CONJ	
rái.	*rai*	rai	NMSNN	13	7	OBL	LOC

7.3. *NBKM 1069* (Beljovo, Panagjurište area, 1776), 137r

такамꙗе било́ писánw на wногѡзи члѡвéка 'thus it was ordained for that man'

text	**diplomatic**	**lemma**	**ME tag**		**UD tag**			
taka+	*taka*	taka	R		1	5	ADVMOD	
mu+	*mu*	toi	PP3MSD		2	8	EXPL	
e	*e*	e	VAIP3SI		3	4	AUX	PRF
biló+	*bilo*	sъm	VAP--SI		4	5	AUX	PASS
pisánw	*pisano*	pisan	VMPA-SI	ANSNN	5	0	ROOT	
na+	*na*	na	SA		6	8	CASE	
wno_gw´zi	*onogozi*	onzi	PD-MSG		7	8	DET	EXT
člw´véka	*človeka*	človek	NMSAY		8	5	OBL	IOBJ

7.4. *PPS* (Mokreš, Vidin-Lom area, 1796), 75v

ѩwстави бащꙗ имайкꙗ ѩдрꙗги сро́дници све що бехꙗ ꙗми́рь 'and she left her father and mother and other relatives, everybody in the world'

text	**diplomatic**	**lemma**	**PoS tag**	**UD tag**		
ʾi+	*i*	i	C	1	2	CC
wstavi	*ostavi*	ostavja	VMIA3SE	2	0	ROOT
baštu	*baštu*	bašta	NFSAY	3	2	OBJ
i+	*i*	i	C	4	5	CC
maĭku	*maiku*	maika	NFSAY	5	3	CONJ
ʾi+	*i*	i	C	6	8	CC
drugi	*drugi*	drug	AMPNN	7	8	AMOD
sródnici	*srodnici*	srodnik	NMPNY	8	5	CONJ
sve	*sve*	vse	NNSNN	9	8	APPOS
što	*što*	što	PQ	10	11	MARK
bexu	*bexu*	sъm	VMII3PI	11	9	ACL
u+	*u*	u	SG	12	13	CASE
mírь	*mirъ*	mir	NMSNN	13	11	OBL	LOC

7.5. *Berl.d.* (Moesian area, 1803), 179v

а̀мико́йщѐ дамо́же да иска́же не́йните по́честй, ра́боти ѝ и́чюдесата 'but who could retell her virtues, works and wonders?'

text	**diplomatic**	**lemma**	**PoS_tag**	**UD tag**			
ami+	*ami*	ami	C	1	7	CC	
kóĭ+	*koi*	koi	PQ-MSN	2	7	NSUBJ	
štè	*šte*	šta	VAIP3SI	3	5	AUX	FUT
da+	*da*	da	C	4	3	FIXED	
móže	*može*	moga	VAIP3SI	5	7	AUX	
da	*da*	da	C	6	5	FIXED	INF
iskáže	*iskaže*	iskaža	VMIP3SE	7	0	ROOT	
néĭni+	*neini*	nein	AFPNN	8	10	AMOD	POSS
te	*te*	тъ	PD-FPN	9	8	DET	P_ADJ
póčestĭ,	*počesti*	počest	NFPNN	10	7	OBJ	
ráboti	*raboti*	rabota	NFPNN	11	10	CONJ	
ì	*i*	tja	PP3FSD	12	11	NMOD	POSS
i+	*i*	i	C	13	14	CC	
čjudesa+	*čjudesa*	čjudo	NNPNN	14	11	CONJ	
ta	*ta*	тъ	PD-NSN	15	14	DET	P_NOM

7.6. *NBKM 1064* (Sliven, Subbalkan area, 1820s), 40r

а̀мѝ дасα ѵ́стάвιм ѵ́т тѿ̀н мпрáтїα мóя χριστéνι 'but let us abandon this, my Christian brothers'

text	**diplomatic**	**lemma**	**PoS tag**	**UD tag**			
amì	*ami*	ami	C	1	4	CC	
da+	*da*	da	C	2	4	AUX	OPT
sà	*sa*	se	PX---A	3	4	EXPL	
ustávım	*ustavim*	ostavja	VMIP1PE	4	0	ROOT	
ut	*ut*	ot	SG	5	6	CASE	
túi	*tui*	тъi	R	6	4	ADVMOD	
mprátïa	*bratia*	brat	NMPNY	7	4	VOCATIVE	
mói	*moi*	moi	AMPNN	8	7	AMOD	
xrısténi.	*xristeni*	xristianin	NMPNY	9	7	APPOS	

7.7. *NBKM 728* (Tetovo, West Macedonia, 1830s), 8r

копаїкї гробо наїдоа тѣло неїзгнїено 'as they dug, they found an undecayed body'

text	**diplomatic**	**lemma**	**PoS tag**		**UD tag**			
kopaïkï	*kopaiki*	kopaja	VMPP-PI		1	3	ADVCL	
grobo	*grobo*	grob	NMSON		2	1	OBJ	P_NOM
naïdoa	*naidoa*	naida	VMII3PE		3	0	ROOT	
tělo	*tělo*	tělo	NNSNN		4	3	OBJ	
ne+	*ne*	ne	QZ		5	6	AMOD	
ïzgnïeno	*izgnieno*	izgnija	ANSNN	VMPP-SI	6	4	AMOD	

8. Character Sets

8.1. Cyrillic

А	a	Є	e	Ї	ї	О	o	У	u	Ш	š	Ꙗ	ja	Ъ	ě	Ѵ	v	ʻ	ˈ
Б	b	Ѕ	ź	К	k	П	p	ф	f	Щ	št	Ѥ	je	Й	ĭ	Ѱ	ψ		
В	v	Ж	ž	Л	l	Р	r	Х	x	Ѣ	ъ	Ю	ju	І	ı	Ѯ	ѯ		
Г	g	Ꙁ	z	М	m	С	s	Ч	č	Ы	y	Ѧ	ę	Ѫ	ų	Ѡ	w		
Д	d	Н	i	Ν	n	Т	t	Ц	c	Ь	ь	Ѫ	ǫ	Ѯ	ѳ	Ѹ	ou		

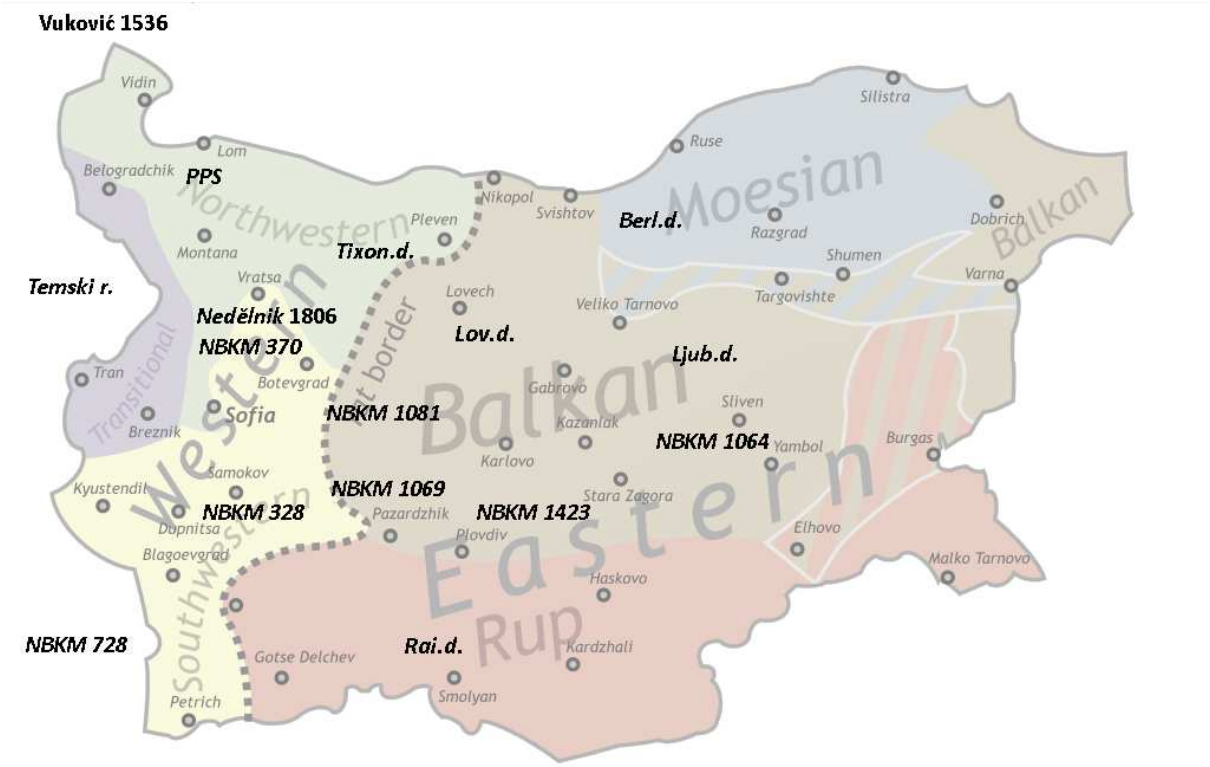
The above set is based on the Resava orthography. It may be accustomed for an individual text. The superscripted letters are written in line with the rest of the word: e.g. the common digraph **ѿ** is written as *wt*. The sequence **шт** (adopted in literary Macedonian) is transcribed as *śt*. The accent (ˊ, ˋ, ^) is marked on the letter, if the combination can be represented by a single UTF character (e.g. *á*). Otherwise, it is written after the vowel (e.g. *ę́*). Double gravis accent (*kendima*), whose function is ambiguous, is marked always after the vowel (e.g. *y̏*). In the same way we mark breves over other sequence-final vowels than й (e.g. *ę̆*). Spirits (ʻ, ʾ) aren't marked in the corpus, if not defined otherwise. Broad (ѥ, ѻ, ѽ) and space-saving (ҁ) variants of certain letters (excl. *paerčik*) are not distinguished.

8.2. Greek

α	a	**ε**	e	**κ**	k	**ο**	o	**ѕ**	s	**φ**	f	**ια**	ja
β	v	**ζ**	z	**λ**	l	**π**	p	**ς**	śt	**χ**	x	**ιѯ**	ju
ϐ	v	**η**	i	**μ**	m	**ϖ**	p	**τ**	t	**ν**	v	**ιω**	jw
γ	g	**ι**	ı	**ν**	n	**ρ**	r	**ҁ**	t	**ψ**	ψ	**τζ**	tz
δ	d	**ϑ**	ѳ	**ξ**	ѯ	**σ**	s	**ѹ**	u	**ω**	w		

Manuscript *NBKM 1064* uses a variant of the modern Greek script based on the Garamond typeface, which has been taught at schools in Southern Bulgaria in the 19th century. Other sources using the script, which are considered for inclusion in the corpus, are the *damaskini NBKM 345* from 1752 and *NBIV 600* from 1860, which both represent dialectal varieties of the Rup area. The above mentioned transliteration methods are also followed in this document.

9. Maps



9.1. Approximate localization of sources on the map of dialects in Bulgaria
source - https://en.wikipedia.org/wiki/Bulgarian_dialects

Literature

Berl.d. - [*Berlin damaskin*], Cracow: Library of Jagellonian University, sign. Slav. fol. 36.

Cod.Zogr. - *Codex Zographensis.* On the basis of the edition by V. Jagić, *Quattuor evangeliorum codex glagoliticus olim Zographensis nunc Petropolitanus, characteribus cyrillicis transcriptum notis criticis prolegomenis appendicibus auctum*, Berolini/Sanktpeterburg, 1879. Available online - link (accessed 08.10.2020)

Dobr.ev. - Цоневъ, Б., *Добрѣйшово четвероевангеле: срѣднобългарски паметникъ отъ XIII вѣкъ*, *Български старини кн. I*, София: Министерството на народната просвѣта, 1906.

Tixon.d./Demina 1972 - Демина Е. И., *Тихонравовский дамаскин: болгарский памятник XVII в.: исследования и текст, Том II*, София: БАН, 1972.

Ljub.d. - [*Ljubljana damaskin*], *Zbornik "Damaskin"*, *bolgarski*, Ljubljana: National and University Library of Slovenia, sign. NUK Cod. Kop. 21. Available online - link (accessed 19.8.2020).

NBIV 600 - *Райковски дамаскин 160(600)*, Plovdiv: National Library of Bulgaria, sign. НБИВ 600. Available online - link (accessed 19.8.2020).

NBKM 328 - *Дамаскин от 1750 г.*, Sofia: National Library of Bulgaria, sign. НБКМ 328.

NBKM 345 - Kail.G., *Zur Sprache der bulgarischen Handschrift "Damaskin von Pazardžik" von 1753*, Wien: Universität Wien (Diplomarbeit), 2013.

NBKM 370 - *Пайсиева История славено-болгарская*, *препис (Еленски) от 1784 г.*, Sofia: National Library of Bulgaria, sign. НБКМ 370.

NBKM 667 - Цонев Б., *Опис на славянските ръкописи в софийската Народна библиотека*, *Том II*, София: Народна библиотека, 1923, 180-181.

NBKM 728 - *Откъслек от простонароден дамаскин, век XIX*, Sofia: National Library of Bulgaria, sign. НБКМ 728.
NBKM 1064 - *Дамаскин, от края на XVIII или началото на XIX в.*, Sofia: National Library of Bulgaria, sign. НБКМ 1064.
NBKM 1069 - *Дамаскин (Бельовски), от 1776 г.*, Sofia: National Library of Bulgaria, sign. НБКМ 1069.
NBKM 1081 - *Сборник от слова и поучения, от началото на XIX в. (1821 г.)*, Sofia: National Library of Bulgaria, sign. НБКМ 1081.
NBKM 1423 - Sofia: National Library of Bulgaria, sign. НБКМ 1423.
Nedělnik 1806 - Софроний еп. Врачански, *Кириакодромион сиреч Неделник - Поучение*, Рымницу: еп. Нектарий, 1806. Available online - link (accessed 19.8.2020).
Nedělnik 1856 - Софроний еп. Врачански, *Евангелие поучително*, пр. Т. Т. Хрулюв, Нов Сад: И. Стоянов, 1856. Available online - link (accessed 19.8.2020).
PPS - *Поп-Пунчов сборник от 1796 год.*, Sofia: National Library of Bulgaria, sign. НБКМ 693. Available online - link (accessed 19.8.2020).
Temski r./Vasilev 1986 - Василев В. П., Темският ръкопис - български езиков паметник от 1764 г., *Старобългаристика*, 1986, кн. 1, 49-72.
Vuković 1536 - [*Zbornik za putnike*. Venecia: Božidar Vuković, 1536.] Available online - link (accessed 19.8.2020).

Bulgarian Etymological Dictionary - Георгиев, В. И. (ед. 1972-2006), *Български етимологичен речник*, *Том I-V*, София: БАН, 1972-1996. Тодоров, Т. А. (ед., 2002-2010), *Том VI-VII*, София: Марин Дринов, 2002-2010.
Cejtlin 1996 - Цейтлин, Р. М. и др., *Старославянский словарь (по рукописям X-XI веков)*, Москва: Русский язык, 1994.
Demina 1968 - Демина Е. И., *Тихонравовский дамаскин: болгарский памятник XVII в.: исследования и текст*, *Том I*, София: БАН, 1968.
Demina et al. 2012 - Дьомина Е. И., Блажева Р., Клепикова Г., Кочева Е., Лилова Р., Манолева А., Мичева В., Петкова А., Сеизова С., Цибранска М., Шаламов Б., *Речник на книжовния български език на народна основа от XVII век (върьу текст на Тихонравовия дамаскин)*, София: "Валентин Траянов", 2012.
Lunt 2001 - Lunt H. G., *Old Church Slavonic Grammar, 7th edition*, Berlin: Mouton de Gruyter, 2001.
Maslov 1981 - Маслов Ю. С., *Грамматика болгарского языка*, Москва: Высшая школа, 1981.
Mladenov 1963 - Младенов М. Сл., "Членувани прилагателни форми на *-ий* в североизточните български говори", *Български език*, 1963, кн. 4-5, 404-410.
Mladenova 2007 - Младенова О. М., "Отново за локализацията на първоначалния новобългарски дамаскинов превод", *Това чудо: езикът! Изследвания в чест на проф. д-р Живко Бояджиев*, София: "Св. Климент Охридски", 2007, 309-316.
Miklosich 1865 - Miklosich Fr., *Lexicon Palaeoslovenico-Graeco-Latinum*, Vindobonae: Guilelmus Braumueller, 1865.
Mirčev 1978 - Мирчев К., *Историческа граматика на българския език*, София: Наука и изкуство, 1978.
Stojkov 2002 - Стойков Ст., *Българската диалектология*, София: Акад. изд. "Проф. Марин Дринов", 2002.
Velčeva 2001 - Велчева Б., "Дамаскините от XVII век и началото на новобългарския книжовен език", *Старобългаристика*, 2001, кн. 4, 64-81.